Introduction:

The paper highlights the shortcomings of traditional Retrieval-Augmented Generation (RAG) systems, which often struggle with context when retrieving information from large knowledge bases. These systems lose crucial context by breaking documents into smaller chunks. To solve this, the authors propose Contextual Retrieval, which uses Contextual Embeddings and Contextual BM25 to enhance retrieval accuracy and performance, leading to better and more relevant information generation by RAG systems

Introductory Contextualised Retrieval in RAG System:

The paper introduces Contextual Retrieval as an extension of traditional Retrieval-Augmented Generation (RAG) systems. The aim is to fix some weaknesses in these traditional methods, where important context is often lost when retrieving from large knowledge bases.

Traditional RAG Systems: The Problem:

Traditional RAG systems break documents into blocks and use embeddings to retrieve similar ones based on meaning. However, this can miss critical context, leading to incomplete or incorrect information retrieval.

Contextual Retrieval: The Solution:

Contextual Retrieval improves RAG by adding chunk-specific context before retrieving information. By offering more context around each chunk, retrieval accuracy improves significantly. This involves two main techniques:

- Contextualized Embeddings: Adding more context to chunks so the model understands the wider meaning.
- Contextual BM25: A ranking function that matches exact terms to improve retrieval accuracy for specific phrases.

Combining Embeddings and BM25:

Combining Contextual Embeddings with Contextual BM25 helps reduce failed retrievals. Embeddings capture broad meanings, while BM25 ensures precise term matching. Together, they balance meaning and accuracy.

Implementation by Claude:

The authors used the Claude AI model to generate context for chunks. By running chunks through Claude with a prompt, developers can automatically add precise context, improving retrieval without manual work.

Conclusion:

This paper demonstrates that Contextual Retrieval is a valuable upgrade to RAG systems, balancing context and cost-efficiency. By combining Contextual Embeddings and BM25, with tools like Claude and Prompt Caching, it ensures accurate and relevant information retrieval from large knowledge bases.